An Optimal Trade-off between Content Freshness and Refresh Cost

Yibei Ling♮ *and Jie Mi*+

♮Applied Research Laboratories, Telcordia Technologies, NJ 07960
+Department of Statistics, Florida International University, Miami, FL 33199

2004

Abstract

Caching is an effective mechanism for reducing bandwidth usage and alleviating server load. However, the use of caching entails a compromise between content freshness and refresh cost. An excessive refresh allows a high degree of content freshness at a greater cost of system resource. Conversely, a deficient refresh inhibits content freshness but saves the cost of resource usages. To address the freshness-cost problem, we formulate the refresh scheduling problem with a generic cost model and use this cost model to determine an optimal refresh frequency that gives the best tradeoff between refresh cost and content freshness. We prove the existence and uniqueness of an optimal refresh frequency under the assumptions that the arrival of content update is Poisson and the age-related cost monotonically increases with decreasing freshness. In addition, we provide an analytic comparison of system performance under fixed refresh scheduling and random refresh scheduling, showing that with the same average refresh frequency two refresh schedulings are mathematically equivalent in terms of the long-run average cost.

1 Introduction

The timely information dissemination is the fundamental driving force that spurs ever-growing technology advancement and development. The widespread of Web technologies makes the Internet a de facto channel for mass distribution of information. Nowaday, popular web sites such as *www.cnn.com* and *www.msn.com* can receive ten millions requests per day [1, 10] with normal request rate of 12,000 per minute, and with peak rate of more than 33,000 per minute during breaking news. Such high demands pose a significant overhead on both serving servers and networks surrounding the serving servers [7]. A variety of approaches and system architectures have been introduced to enable efficient content distribution while alleviating system load and bandwidth consumption.

Web servers and browsers (web clients) are the fundamental architectural building blocks in the World Wide Web. A Web client is a requester of data (content) and a Web server is the provider of data (content). A web server manages and provides data source while Web browsers send requests to a Web server for a specific source data by means of URL (uniform resource locator). Upon receipt of a request initiated by a web client, the web server then processes the request and sends a response back to the web client. Figure 1 illustrates the typical data flow between the clients and the server.

The content resources at servers are autonomous: they are updated independently at various rates without pushing updates to the clients. As a result, each client has to poll the remote
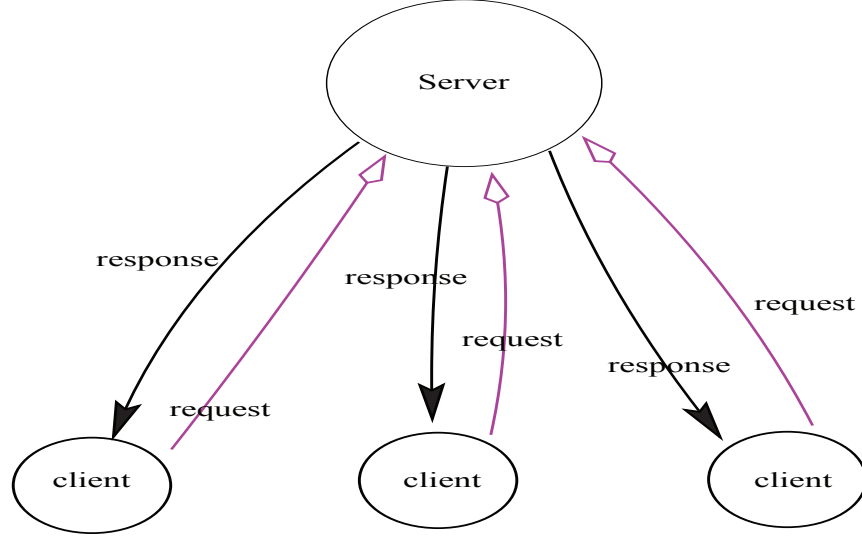
1

Figure 1: Processing flow between clients and a server

resources at a web server periodically in order to detect changes and update its contents. This process is referred to as *refresh synchronization*. The freshness comes at a cost of resource usage; each request initiated by a client incurs certain communication and computation overhead for processing the request. As a result, it could be very costly in terms of overall bandwidth usage and server load when considering million of individual clients.

Caching is an effective means of reducing the system load of servers and bandwidth usage. The idea behind caching is to store recently retrieved copies of remote source somewhere between the clients and the remote web servers. As a result, the request initiated by a web client can be diverted to the cached copies which are much closer to the Web client than the remote web server in terms of network distance.

The caching architecture in a representative enterprise environment is illustrated in Figure 2 wherein a caching server is placed at the enterprise's network entrance to external networks, acting as an intermediary between host computers (clients) inside the enterprise network and the internet. Each host machine is connecting to the enterprise's network backbone, the caching server rests between the enterprise's network backbone and the Internet. Upon receiving a request originated from a host inside the enterprise's network, the caching server checks to see whether the corresponding response has already been cached, if the cached file is present, then the caching server returns the client with the cached copy, saving the client from retrieving the same resource (document) repeatedly from the remote server. The flow is represented by the dotted line 1 in Figure 2. If the cached file is absent, the caching server then forwards the request to a server. After the caching server has received the response from the remote server, it returns the response to the client and locally stores the response for subsequent requests. Its is represented by the dashed line 2 in Figure 2. It is clear that the use of caching server in an enterprise environment enables substantial bandwidth saving and dramatically improves user-perceived response time, because the cached copies are located within the vicinity of the clients in terms of network distance. The performance gain appears to be proportional to the number of users [5, 4, 6].

The performance gains of caching carry the cost of content freshness. The cached copies imme-
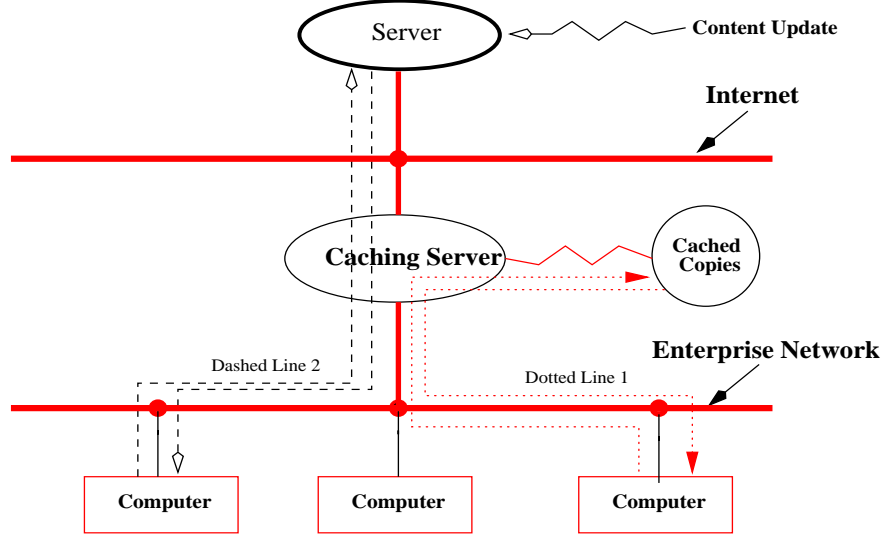
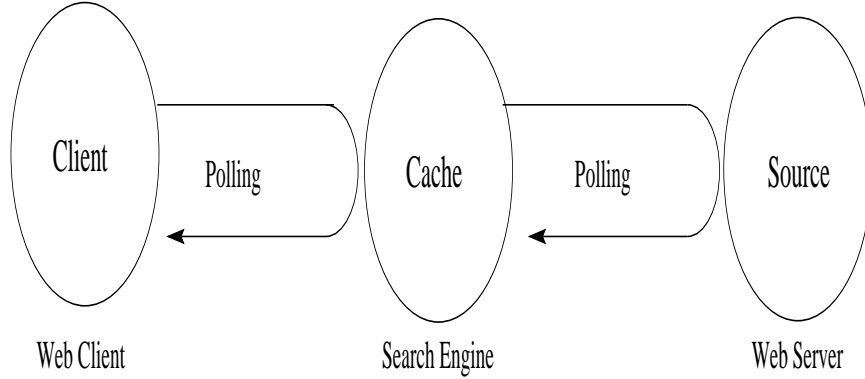Figure 2: Placement of caching server at an enterprise environment

Figure 3: Conceptual Flow Diagram of Crawler-based Search Engine

diately become obsolete as soon as the original copies at the remote server are updated or changed. There is a substantial tradeoff between content freshness and refresh cost: a frequent refresh ensures the freshness of content at high cost of refresh. Conversely, an infrequent refresh inhibits the freshness of content but saves the refresh cost.

The freshness-cost problem also arises in crawler-based search engine applications. Crawler-based search engines such as *google* provide a powerful tool for searching web documents, serving as a "yellow book" on the Web. Figure 3 presents a conceptual flow diagram of crawler-based Web search engines. Periodically, a Web search engine polls Web servers independently, and Web clients poll randomly the Web search engine searching for the directories of Web documents. It is clear the refresh scheduling of the web search engine is independent of the Web access by Web clients. In addition, the content at Web servers may change over time and it is impossible to know a priori exactly the arrival time of content update. To maintain the freshness of content, the Web search engine needs to poll web sites and update its database directory in a frequent fashion. Such a process is known as *Web crawling*. The freshness of content is regarded as one of the important

performance metrics for a crawler-based search engine [8]. Web crawling is a prohibitively expensive computation task when considering an astronomical numbers of constantly changing web pages. It is reported that most search engines refresh their entire directory databases once a month [8].

Cho and Garcia-Molina [2, 3] first present a probability model to study the impact of various refresh (synchronization) policies on the content freshness with an emphasis on synchronization-order policy, under different contexts from this paper. Our study in this paper differs from theirs principally in that we consider the problem of refresh scheduling, with the objective of optimally balancing the tradeoff between content freshness and refresh cost.

The remainder of the paper is organized as follows: In *Section 2* we consider the problem of refresh scheduling involving one cache element. We study and establish a generic cost model that accounts for both the content freshness and the refresh cost and prove that the mathematical equivalence between the refresh schedulings with the fixed interval and the random interval in terms of the overall refresh cost. *Section 3* extends the obtained results into the cases involving more than one cache elements, with an emphasis on the *uniform allocation policy*. *Section 4* concludes the paper.

2   Mathematical Formulation and Main Results

In this section we consider the problem of optimal refresh scheduling involving only one cache element. We begin with an introduction of relevant notions and definitions, followed by a cost analysis of the relationship between the refresh interval and the freshness of content. Finally we identify the optimal refresh frequency that gives the best tradeoff between the content freshness and the refresh cost.

We formally give the notion of the aggregated age function. Our definition is in spirit similar to the one proposed by Cho and Garcia-Molina [3, 2], but differs in the sense that we take the "aggregated effect" of content decay into account. We then introduce the age-related cost function that generalizes the notion of the aggregated age function.

Suppose that the arrival of content update at a server follows the Poisson process with intensity rate of λ. Let $\{X_i, i \geq 1\}$ be the interarrival times of the Poisson process. Define $S_0 = 0$, $S_n = \sum_{i=1}^{n} X_i$, where S_i represents the time of the ith occurrence of content update at the server. Let

$$N(t) = \sup\{n \geq 0 : \ S_n \leq t\}. \tag{1}$$

$N(t)$ is a random variable that represents the number of arrivals of content updates at the server in the time interval $(0, t]$. Two closely related but different notions are given as follows.

Definition 1 *Under refresh frequency of $1/T$, the age of the element e with respect to the ith occurrence of content update S_i at time $t \in [0, T)$ is*

$$Age(e, S_i, t) = (t - S_i)I_{\{t > S_i\}},$$

where $I_{\{t > S_i\}}$ is an indicator function.

$Age(e, S_i, t)$ is a function representing a measure for the content freshness of the element e with respect to the ith occurrence of content update at the server.

Definition 2 *Under refresh frequency of $1/T$, the aggregated age of the element e with respect to the occurrences of content update at time $t \in [0, T)$ is*

$$A(e, t) = \sum_{i=1}^{N(t)} (t - S_i) I_{\{t > S_i\}}.$$

$A(e, t)$ is an aggregated age function that reflects the additive effect of multiple content updates taking place within the interval $[0, t)$. Cho and Garcia-Molina [2, 3] propose the *age* metric as a measure for content freshness by only considering the first occurrence of content update, that is, $Age(e, S_1, t)$. The major difference between ours and the definition by Cho and Garcia-Molina is that we consider the additive property of content freshness with respect to multiple content updates.

Figure 4 is an illustration of the evolution of the functions of the $A(e, t)$, reflecting that the aggregated age of the element e with respect to the occurrences of content update over time.
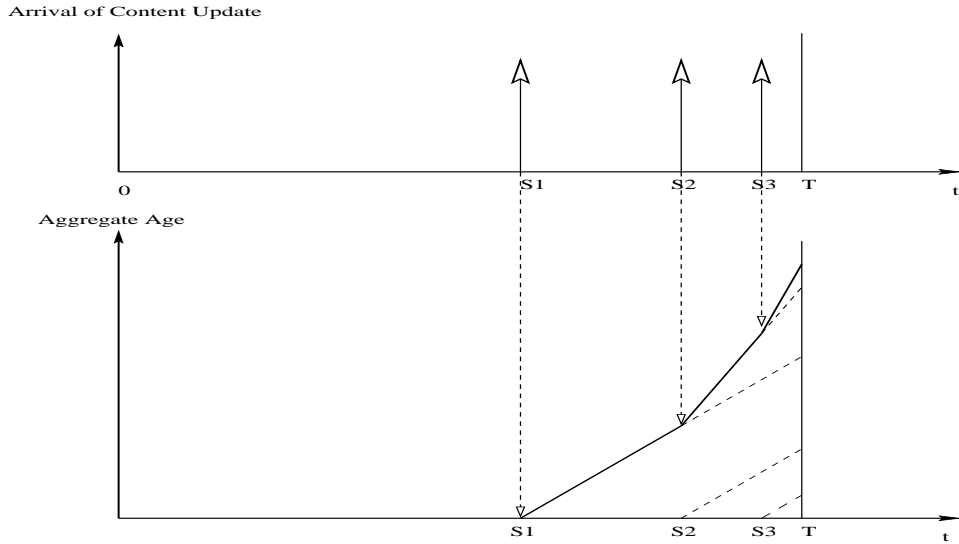

Figure 4: Evolution of Aggregated Age Function of $A(e, t)$

To study the freshness-cost tradeoff, we introduce the age-related cost function denoted by $C_a(x)$ which is a nondecreasing and positive function of the age, where the variable x denotes the age of the element of interest and the subscript a the age of content (see Definition 1). The notion of age-related cost is a generalization of the notion of age. As a result, the age function defined in [2] is a special case of the age-related cost function with $C_a(x) = x$.

The association of cost with freshness in the problem formulation is partly motivated by its market relevance that many crawler-based search engines such as *google*, *Inktomi* and *fast* [8] introduce paid inclusion programs that trade the freshness of content and visibility for a payment. We further assume that the cost of a refresh synchronization C_r, where the subscript r indicates the cost associated with refresh. This cost could be a measure of bandwidth usage or latency or a financial payment, depending on the choice of performance metric. The following theorem shows that under mild conditions the optimal refresh interval, T^*, that minimizes the long-run mean average cost, exists and is unique.

5

Theorem 1 *Suppose that the arrival of update is a Poisson process with intensity rate λ. Let the refresh interval be T, and the cost of a refresh be C_r. Then the long-run mean average cost of caching under the refresh frequency of $1/T$ is given as*

$$C(T) = \frac{C_r}{T} + \frac{\lambda \int_0^T C_a(t)dt}{T}. \tag{2}$$

If the age-related cost function $C_a(t)$ satisfies (1) $C_a(t)$ is differentiable and $C_a'(t) > 0$; (2) $C_a(\infty) = \infty$, then there exists a unique refresh interval, designated T^*, which minimizes the long-run mean average cost function $C(T)$, and T^* is determined by the equation

$$TC_a(T) - \int_0^T C_a(t)dt = \frac{C_r}{\lambda}.$$

Proof For a given interval $(0, t]$, the mean average cost over the interval $(0, t]$ can be written as

$$\frac{E(\text{the random cost in interval } (0, t])}{t}. \tag{3}$$

Hence, the long-run mean average cost under the refresh interval T is

$$C(T) = \lim_{t \to \infty} \frac{E(\text{the random cost in interval } (0, t])}{t} \tag{4}$$

if this limit exists. Denote the random cost on the interval $((k-1)T, kT]$ by $\xi_k(T), k \geq 1$. Due to the properties of stationary and independent increments of the Poisson process [9], $\xi_k(T), k \geq 1$ are *iid*, the long-run mean average cost can be written as

$$C(T) = \lim_{t \to \infty} \frac{E(\lfloor \frac{t}{T} \rfloor)\xi_1(T))}{t} = \frac{E(\xi_1(T))}{T}, \tag{5}$$

where the term $\lfloor x \rfloor$ is the floor function that gives the largest integer less than or equal to x.

The random cost $\xi_1(T)$ on the interval $(0, T]$ consists of a refresh cost and the age-related cost accumulated over the interval $(0, T]$. For the arrival of the nth content update at time $S_n \leq T$, the age-related cost, $C_a(T - S_n)$, is a function of the age $T - S_n$. The aggregated age-related cost in the interval $(0, T]$ is thus expressed as

$$\sum_{n=1}^{N(T)} (C_a(T - S_n)) \, I_{\{N(T)>0\}}.$$

Hence, the random cost $\xi_1(T)$ on the interval $(0, T]$ is given as

$$\xi_1(T) = C_r + \sum_{n=1}^{N(T)} (C_a(T - S_n)) \, I_{\{N(T)>0\}}. \tag{6}$$

Note that

$$\sum_{n=1}^{N(T)} C_a(T - S_n)I_{\{N(T)>0\}} = \sum_{n=1}^{\infty} C_a(T - S_n)I_{\{n \leq N(T)\}} = \sum_{n=1}^{\infty} C_a(T - S_n)I_{\{S_n \leq T\}} \tag{7}$$

and

$$E\left(C_a(T - S_n)I_{\{S_n \le T\}}\right) = \int\limits_0^T C_a(T - t)f_n(t)dt, \tag{8}$$

where $f_n(t)$ is the probability density function of S_n, and follows the gamma distribution as below

$$f_n(t) = \frac{\lambda^n}{(n-1)!}t^{n-1}e^{-\lambda t}, \ t > 0. \tag{9}$$

Substituting Eq(9) into Eq(8) we obtain

$$E\left(C_a(T - S_n)I_{\{S_n \le T\}}\right) = \int_0^T C_a(T - t)\frac{\lambda^n}{(n-1)!}t^{n-1}e^{-\lambda t}dt. \tag{10}$$

We then have

$$\begin{aligned}
E\left(\sum_{n=1}^{N(T)} C_a(T - S_n)I_{\{N(T)>0\}}\right) &= \sum_{n=1}^{\infty}\int_0^T C_a(T - t)\frac{\lambda^n}{(n-1)!}t^{n-1}e^{-\lambda t}dt \\
&= \int_0^T C_a(T - t)\lambda e^{-\lambda t}\left(\sum_{n=1}^{\infty}\frac{(\lambda t)^{n-1}}{(n-1)!}\right)dt \\
&= \lambda\int_0^T C_a(T - t)dt = \lambda\int_0^T C_a(t)dt. \tag{11}
\end{aligned}$$

Therefore, by Eqs(5), (6), and (11), the long-run mean average cost can be expressed as

$$C(T) = \frac{E(\xi_1(T))}{T} = \frac{C_r}{T} + \frac{\lambda\int_0^T C_a(t)dt}{T}.$$

The derivative of $C(T)$ is then given as

$$C'(T) = -\frac{C_r}{T^2} + \frac{\lambda C_a(T)}{T} - \frac{\lambda\int_0^T C_a(t)dt}{T^2}.$$

Define a function $\varphi(T)$ as

$$\varphi(T) \equiv T^2 C'(T) = -C_r + \lambda T C_a(T) - \lambda\int_0^T C_a(t)dt \tag{12}$$

Observe that $C'(T)$ and $\varphi(T)$ have the same sign. It can also be verified that the

$$\varphi'(T) = \lambda T C_a'(T) > 0, \ \ \forall \ T > 0$$

since $C_a'(t) > 0, \forall t > 0$. This shows that the function $\varphi(T)$ strictly increases in $T > 0$. Observe that $\varphi(0) = -C_r < 0$. Moreover, for any fixed $\epsilon > 0$, if $T > \epsilon$, then

$$\begin{aligned}
\varphi(T) &= -C_r + \lambda\int_0^T (C_a(T) - C_a(t))\,dt \ge -C_r + \lambda\int_0^{\epsilon}(C_a(T) - C_a(t))dt \\
&\ge -C_r + \lambda(C_a(T) - C_a(\epsilon))\epsilon \tag{13}
\end{aligned}$$

Hence $\varphi(\infty) \equiv \lim\limits_{T \to \infty} \varphi(T) = \infty$ since $C_a(\infty) = \infty$. Considering the facts that $\varphi(0) < 0$, $\varphi(\infty) = \infty$ and $\varphi(T)$ is strictly increasing in $T > 0$, we conclude that there must exist a unique $0 < T^* < \infty$ such that

$$C'(T) = \begin{cases} < 0, & \text{if } 0 < T < T^* \\ = 0, & \text{if } T = T^* \\ > 0, & \text{if } T > T^*. \end{cases}$$

Therefore, T^* minimizes $C(T)$, or

$$T^* = \arg\left(\min_{T>0} C(T)\right).$$

∎

Corollary 1 *Suppose that the age-related cost function $C_a(t) = Ct$, where $C > 0$ is the proportionality constant. Then the long-run mean average cost function $C(T)$ is given as*

$$C(T) = \frac{C_r}{T} + \frac{C\lambda T}{2} \tag{14}$$

and is minimized at

$$T^* = \sqrt{\frac{2C_r}{\lambda C}}. \tag{15}$$

Also, $C(T^*) = \sqrt{2\lambda C C_r}$.

Proof Obviously $C_a(t) = Ct$ satisfies the condition in Theorem 1. Thus, $T^* \equiv \arg(\min\limits_{T>0} C(T))$ exists and is uniquely determined by the equation $C'(T) = 0$ where

$$C'(T) = -\frac{C_r}{T^2} + \frac{\lambda C}{2}$$

and the unique solution to $C'(T) = 0$ is given by

$$T^* = \sqrt{\frac{2C_r}{\lambda C}}$$

∎

.

From Corollary 1, it is shown that the optimal refresh interval decreases with increasing rate of the content update when the age function is linear. This result, however, is also true in general as shown in the following theorem.

Theorem 2 *Under the same conditions of Theorem 1, the optimal refresh interval T^* strictly decreases in $\lambda > 0$.*

Proof Clearly T^* is a function of λ. For the sake of simplicity, we will not express this dependence explicitly. For the same reason we will suppress the symbol $*$ and just use T to denote the optimal

refresh interval in the following derivation. From Theorem 1, the optimal refresh interval satisfies $\varphi(T) = 0$, or

$$-C_r + \lambda T C_a(T) - \lambda \int_0^T C_a(t)dt = 0. \tag{16}$$

Taking derivative with respect to λ in Eq(16), we obtain

$$TC_a(T) + \lambda T C_a'(T)\frac{dT}{d\lambda} - \int_0^T C_a(t)dt = 0. \tag{17}$$

From Eq(17) it follows that

$$\begin{aligned} \lambda T C_a'(T)\frac{dT}{d\lambda} &= \int_0^T C_a(t)dt - TC_a(T) \\ &= \int_0^T \left(C_a(t) - C_a(T) \right) dt < 0. \end{aligned} \tag{18}$$

Eq(18) immediately implies

$$\frac{dT^*(\lambda)}{d\lambda} < 0$$

since $C_a'(T) > 0$. This proves that $T^* = T^*(\lambda)$ is a strictly decreasing function of λ. ∎

Theorem 3 *Under the same conditions of Theorem 1, the fixed optimal refresh policy T^* strictly increases in $C_r > 0$.*

The proof is simple and straightforward, therefore is omitted.

Remark 1: Theorems 2, 3 demonstrate that the impact of the arrival rate of content update as well as the refresh overhead C_r on the optimal refresh frequency. A high frequency of content update and cheap refresh cost exact a high frequency of refresh in order to maintain a certain level of content freshness. Conversely, an infrequent content update and expensive refresh cost require refresh be performed at low frequency in order to save bandwidth usage and mitigate server load. This analytical results obtained above not only agree well with our intuition, but also provide us with a quantitative connection between optimal refresh interval and arrival rate of update.

Theorem 4 *Suppose that two age-related cost functions $C_{a1}(t)$ and $C_{a2}(t)$ satisfy the conditions in Theorem 1. Let T_1^* and T_2^* be the optimal refresh intervals associated with $C_{a1}(t)$ and $C_{a2}(t)$, respectively. For given C_r and λ, if $0 \le \Delta(t) \equiv C_{a2}(t) - C_{a1}(t)$ is nondecreasing in $t \ge 0$, then $T_1^* \ge T_2^*$.*

Proof According to the proof of Theorem 1, it is shown that

$$\varphi(T) \equiv -C_r + \lambda T C_{a1}(T) - \lambda \int_0^T C_{a1}(t)dt \tag{19}$$

strictly increases in T, and

$$\varphi(T_1^*) = -C_r + \lambda T_1^* C_{a1}(T_1^*) - \lambda \int_0^{T_1^*} C_{a1}(t)dt = 0. \tag{20}$$

9

Suppose that contrary to the claimed result, i.e., $T_1^* < T_2^*$. From the monotonicity of $\varphi(T)$ in Eq(12) and $\varphi(T_1^*) = 0$, we have $\varphi(T_2^*) > 0$. That is

$$-C_r + \lambda T_2^* C_{a1}(T_2^*) - \lambda \int_0^{T_2^*} C_{a1}(t)dt > 0. \tag{21}$$

Note that

$$
\begin{aligned}
&-C_r + \lambda T_2^* C_{a2}(T_2^*) - \lambda \int_0^{T_2^*} C_{a2}(t)dt \\
=~& -C_r + \lambda T_2^* C_{a1}(T_2^*) + \lambda T_2^* \Delta(T_2^*) - \lambda \int_0^{T_2^*} C_{a1}(t)dt + \lambda \int_0^{T_2^*} \Delta(t)dt \\
=~& \left\{ -C_r + \lambda T_2^* C_{a1}(T_2^*) - \lambda \int_0^{T_2^*} C_{a1}(t)dt \right\} + \lambda \left\{ T_2^* \Delta(T_2^*) - \int_0^{T_2^*} \Delta(t)dt \right\} \\
=~& \varphi(T_2^*) + \lambda \int_0^{T_2^*} \left(\Delta(T_2^*) - \Delta(t) \right) dt \\
\geq~& \varphi(T_2^*) > 0, \tag{22}
\end{aligned}
$$

where Eq(22) comes directly from the assumption that $\Delta(t)$ is nondecreasing in $t \geq 0$. However, as the optimal refresh policy associated with $C_{a2}(t)$, it must be true that

$$-C_r + \lambda T_2^* C_{a2}(T_2^*) - \lambda \int_0^{T_2^*} C_{a2}(t)dt = 0$$

which contradicts with the inequality (22). Therefore, we must have $T_1^* \geq T_2^*$. ∎

In the previous discussion we assume that refresh interval has a fixed length of T. This assumption is somewhat unrealistic in practice. In the following, we study the impact of random refresh scheduling on the long-run mean average cost.

Let $\{Y_i, i \geq 1\}$ be a sequence of *iid* random variables defined on $(0, \infty)$ following certain distribution H. The sequence $\{Y_i, i \geq 1\}$ representing the interarrival times under a random refresh scheduling is assumed to be independent of the Poisson arrival of content update at a server. It is obvious that a fixed refresh policy is only a special case of a random refresh policy.

Let $\mathcal{H}$ be the family of all distribution functions on $(0, \infty)$ with finite first moment. Namely,

$$\mathcal{H} = \left\{ H \colon H \text{ is a CDF on } (0, \infty) \text{ and } \int_0^\infty \bar{H}(t)dt < \infty \right\}$$

where $\bar{H}(t) \equiv 1 - H(t), ~\forall t \geq 0$.

Theorem 5 *Let $C_H(\cdot)$ denote the long-run mean average cost under a random refresh interval Y characterized by certain distribution function $H \in \mathcal{H}$, and $C(T)$ denote the long-run mean average cost under fixed refresh scheduling with interval T, then*

$$\min_{H \in \mathcal{H}} C_H = \min_{T > 0} C(T).$$

Proof Since the sequence $\{Y_i, i \geq 1\}$ of interarrival refresh times is independent of the Poisson arrival of content update, it is easy to see that the random costs over the intervals $(0, Y_1], (Y_1, Y_1 + Y_2], \ldots$ are *iid*. Using the same line as the proof of Theorem 1, the long-run mean average cost is expressed as

$$C_H = \frac{E(\text{random age-related cost over } (0, Y])}{E(Y)},$$

where Y has distribution H. Let $\xi(Y)$ be the random cost in the cycle $(0, Y]$, we have

$$
\begin{aligned}
E(\xi(Y)) &= E\{E[\xi(Y)|Y]\} \\
&= \int_0^\infty E(C_r + \sum_{n=1}^{N(y)} C_a(y - S_n) I_{\{N(y)>0\}}) dH(y),
\end{aligned}
\tag{23}
$$

where $S_n = \sum_{i=1}^n X_i$ denotes the time of the nth arrival of content update at the server. Due to the independence of $\{X_i, i \geq 1\}$ and $\{Y_i, i \geq 1\}$, from Eq(23) we further obtain

$$
\begin{aligned}
E(\xi(Y)) &= \int_0^\infty (C_r + \int_0^y \lambda C_a(t) dt) dH(y) \\
&= C_r + \int_0^\infty \left(\int_0^y \lambda C_a(t) dt \right) dH(y).
\end{aligned}
$$

Therefore, the long-run expected average cost is expressed as

$$C_H = \frac{C_r}{E(Y)} + \frac{\int_0^\infty (\int_0^y \lambda C_a(t) dt) dH(y)}{E(Y)}.
\tag{24}$$

It is straightforward that

$$\min_{H \in \mathcal{H}} C_H \leq \min_{T > 0} C(T)
\tag{25}$$

since a fixed refresh scheduling with the interval $T > 0$ is a degenerate random refresh scheduling when $P(Y = T) = 1$.

Now for any given random refresh scheduling Y with distribution $H \in \mathcal{H}$. Eq(24) can be reexpressed as

$$
\begin{aligned}
C_H &= \frac{C_r}{E(Y)} + \frac{\int_0^\infty (\int_t^\infty \lambda C_a(t) dH(y)) dt}{E(Y)} \\
&= \frac{C_r}{E(Y)} + \frac{\lambda \int_0^\infty C_a(t) \bar{H}(t) dt}{E(Y)}.
\end{aligned}
\tag{26}
$$

If we choose $T = E(Y) \equiv \mu$, meaning that the fixed refresh interval T equals to the mean value of the random refresh interval Y, then according to Eq(2), we have

$$C(T) = \frac{C_r}{\mu} + \frac{\lambda \int_0^\mu C_a(t) dt}{\mu}.
\tag{27}$$

Subtracting Eq(27) from Eq(26), we obtain

$$C_H - C(T) = \frac{\lambda}{\mu} \left\{ \int_0^\infty C_a(t) \bar{H}(t) dt - \int_0^\mu C_a(t) dt \right\}$$

11

$$
\begin{aligned}
&= \frac{\lambda}{\mu}\left\{\int_0^\mu C_a(t)\bar{H}(t)dt + \int_\mu^\infty C_a(t)\bar{H}(t)dt - \int_0^\mu C_a(t)dt\right\} \\
&= \frac{\lambda}{\mu}\left\{\int_\mu^\infty C_a(t)\bar{H}(t)dt - \int_0^\mu C_a(t)H(t)dt\right\} \\
&\geq \frac{\lambda}{\mu}\left\{C_a(\mu)\int_\mu^\infty \bar{H}(t)dt - C_a(\mu)\int_0^\mu H(t)dt\right\} \\
&= \frac{\lambda C_a(\mu)}{\mu}\left\{\int_\mu^\infty \bar{H}(t)dt - \int_0^\mu (1 - \bar{H}(t))dt\right\} \\
&= \frac{\lambda C_a(\mu)}{\mu}\left\{\int_0^\infty \bar{H}(t)dt - \mu\right\} = 0
\end{aligned}
\tag{28}
$$

thus we have

$$
C_H \geq C(T) \tag{29}
$$

Combining Eq(25) and Eq(29), we obtain

$$
\min_{H \in \mathcal{H}} C_H = \min_{T > 0} C(T)
$$

∎

Such a mathematical equivalence between fixed refresh and random refresh schedulings offers a theoretical justification for the flexibility that is needed in real application environments. Refresh scheduling with random variable Y over a finite interval (a, b) is more flexible than fixed refresh scheduling. Let $T > 0$ be a given fixed refresh interval and Y be a positive random interval following a distribution on interval (a, b) with mean $E(Y) = T$. From Eq(29) it can be seen that $C_H \geq C(T)$ and the equality holds if and only if Y is a degenerate random variable when $P(Y = T) = 1$. Thus, for any given $T > 0, \delta > 0$ satisfying $T - \delta > 0$, if we denote $\mathcal{H}^* = \{H : H \text{ is a CDF on } (T - \delta, T + \delta) \text{ and } \int_{T-\delta}^{T+\delta} t dH(t) = T\}$, then we have $C_H > C(T)$ for any non-degenerate $H \in \mathcal{H}^*$. In particular, if $U(\delta)$ is the uniform distribution on $(T - \delta, T + \delta)$, then it is true that $C_{U(\delta)} > C(T)$. It is also easy to show that $\lim_{\delta \to 0} C_{U(\delta)} = C(T)$.

3 Synchronizing Multiple Content Elements

In the section we will extend this approach to the case that involves more than one elements with different content update rates.

Consider a server containing content elements $S = \{e_1, \cdots, e_M\}$. Each element e_i is being updated according to a Poisson process with intensity λ_i ($1 \leq i \leq M$). It is theoretically appealing to synchronize each content element e_i using refresh interval T_i, which is referred to as *non-uniform allocation policy* [2, 3]. However, *uniform allocation policy* [2, 3], i.e., synchronizing all content elements by using the same refresh interval T, is practically preferable by amortized cost analysis. The underlying reason is that each refresh requires a connection establishment along with bandwidth usage and processing overhead. A connection overhead (latency) denoted as C_{conn} is considered as the dominating factor in determining refresh cost C_r. The advantage of the *uniform allocation policy* over the *non-uniform allocation policy* is its ability to share its connection cost over the number of the content elements. As a result, the amortized connection for the *uniform allocation policy* is calculated as C_{conn}/M, which is much cheaper than that of the *non-uniform allocation policy*, C_{conn}.

Let λ_i be the intensity rate of Poisson process describing the content update of the cache element e_i, $1 \le i \le M$. From Theorem 1, the long-run mean average cost of caching over the entire S is

$$C(T) = \sum_{i=1}^{M} \left[\frac{C_r(\lambda_i)}{T} + \frac{\lambda_i \int_0^T C_a(t; \lambda_i)dt}{T} \right],$$

where $C_r(\lambda_i)$ is the cost associated with a refresh of element e_i, and $C_a(t; \lambda_i)$ is the age-related cost function associated with element e_i, if the *uniform allocation policy* T is applied. Then $C(T)$ can be rewritten as

$$C(T) = M \left[\frac{\bar{C}_r}{T} + \frac{\int_0^T \bar{C}_a(t)dt}{T} \right], \tag{30}$$

where

$$\bar{C}_r = \frac{\sum_{i=1}^{M} C_r(\lambda_i)}{M} \tag{31}$$

and

$$\bar{C}_a(t) = \frac{\sum_{i=1}^{M} \lambda_i C_a(t; \lambda_i)}{M}. \tag{32}$$

If the size of the set of cache elements, M, is sufficiently large, then it is convenient to describe all λ_i by a probability density function $g(\lambda)$ on $(0, \infty)$. With this assumption, the long-run mean average cost of caching the entire set of elements S is given as

$$C(T) = M \left[\frac{\bar{C}_r}{T} + \frac{\int_0^T \bar{C}_a(t)dt}{T} \right], \tag{33}$$

where

$$\bar{C}_r = \int_0^\infty C_r(\lambda)g(\lambda)\, d\lambda$$

and

$$\bar{C}_a(t) = \int_0^\infty \lambda C_a(t; \lambda)g(\lambda)\, d\lambda.$$

Combining Eq(30) and Eq(33), we see that minimizing $C(T)$ is equivalent to minimizing

$$\frac{\bar{C}_r}{T} + \frac{\int_0^T \bar{C}_a(t)dt}{T}. \tag{34}$$

Since the form of Eq(34) is the same as Eq(2), we immediately obtain the following result.

Theorem 6 *Suppose that the arrival of update of each content e is a Poisson process with intensity rate λ, $C_r(\lambda)$ is the cost associated with a refresh, and the refresh interval is T. Then the long-run mean average cost of caching S is given as*

$$C(T) = M \left[\frac{\bar{C}_r}{T} + \frac{\int_0^T \bar{C}_a(t)dt}{T} \right],$$

where

$$\bar{C}_r = \int_0^\infty C_r(\lambda)\, dG(\lambda),$$

$$\bar{C}_a(t) = \int_0^\infty \lambda C_a(t; \lambda)\, dG(\lambda),$$

and $G(\lambda)$ is a distribution function on $(0, \infty)$ satisfying $\int_0^\infty \lambda C_a(t; \lambda) \, dG(\lambda) < \infty$ for any $t > 0$. The problem of minimizing $C(T)$ is equivalent to minimizing $\bar{C}(T)$ given as

$$\bar{C}(T) = \frac{\bar{C}_r}{T} + \frac{\int_0^T \bar{C}_a(t) dt}{T}. \tag{35}$$

If the age-related cost function $\bar{C}_a(t)$ satisfies (a) $\bar{C}_a'(t) > 0$; (b) $\bar{C}_a(\infty) = \infty$, then there exists unique refresh interval, designated T^*, which minimizes $C(T)$ and $\bar{C}(T)$. Moreover, T^* is determined by the equation

$$T\bar{C}_a(T) - \int_0^T \bar{C}_a(t) dt = \bar{C}_r.$$

4 Conclusion

There is a fundamental tradeoff between the freshness of content and the overhead of refresh synchronization. An excessive refresh puts a strain on computation resource of network bandwidth and servers, while a deficient refresh sacrifices the required freshness of content. This paper is focused on studying the effect of refresh scheduling on the freshness-cost tradeoff. We formulate the refresh scheduling problem with a generic cost model to capture the relationship between the arrival rate of content update, the freshness of content and the cost of refresh synchronization, and then use this cost model to determine an optimal refresh interval that minimizes the overall cost involved.

Theoretical results obtained in this paper suggest that the optimal refresh frequency should be a function of the arrival rate of content update and the refresh cost, implying that refresh frequency is determined by the arrival rate of content update in order to maintain a certain level of content freshness. Such a viewpoint has been implicitly reflected in the fact that many crawler-based Web search engines like *google*, *Inktomi* and *fast* started developing an automated tool to identify the content change rate at some Web sites, and adapt refresh rate to the actual changing rate of content sources [8]. This paper gives a quantitative analysis of the freshness-cost tradeoff and offers a theoretical guidance in determining the best tradeoff between the freshness of content and the cost of refresh synchronization.

5 Acknowledgement

References

[1] Cnn delivers unprecedented online service. In *http://www.volera.com/corporate/pressroom/casestudies/cnn.ht*

[2] Junghoo Cho and Hector Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *Proceedings of the Twenty-sixth International Conference on Very Large Databases*, 2000.

[3] Junghoo Cho and Hector Garcia-Molina. Synchronizing a database to improve freshness. In *Proceedings of the 2000 ACM International Conference on Management of Data*, June 2000.

[4] Edith Cohen and Haim Kaplan. Age penalty and its effects on cache performance. In *The 3rd USENIX Symposium On Internet Technologies and Systems*, March 2001.

[5] Edith Cohen and Haim Kaplan. Refreshment policies for web content caches. In *Proceedings of Proceedings of the IEEE INFOCOM'01 Conference*, 2001.

[6] Bradley M. Duska, David Marwood, and Michael J. Feeley. The measured access characteristics of world-wide-web client proxy caches. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, December 1997.

[7] Kang-Won Lee, Khalil Amiri, S. Sahu, and Chitra Venkatramani. On the sensitivity of cooperative caching performance to workload and network characteristics. In *Proceedings of the ACM Conference on Measurement and Modeling of Computer SystemS (SIGMETRICS)*, 2002.

[8] Greg Notess. Freshness issue and complexities with web search engines. *OnLine*, 25(6), November 2001.

[9] Sheldon M. Ross. *Stochastic Processes*. John Wiley & Sons, Inc., New York, 1996.

[10] Duane Wessels. *Web Caching*. O'Reilly & Associates, Inc, USA, 2001.